Imperial College
London

7 September 2023

# Combining non-exchangeable stochastic processes and dynamic graph VAEs for anomaly detection in computer networks

Alberto Mazzetto

Supervised by Prof. F. Sanna Passino

Imperial College London

# Table of Contents I

Imperial College London

# Outline

# Introduction. Motivating context

- Scale and complexity of **cybersecurity threats** are raising [19, 20] $\rightarrow$ economic, democratic and stability costs;
- Made it necessary to introduce **statistically principled methods**, other than signature-based methods $\rightarrow$ better generality;
- Main challenges: **fast computations** and **reduced false alarms**;
- Silver bullet: a network intrusion is often characterised by **traversal movement**, likely to create unexpected communication patterns [4, 5];
- The most natural representation of a computer network is as **time-aggregated dynamic graph**:

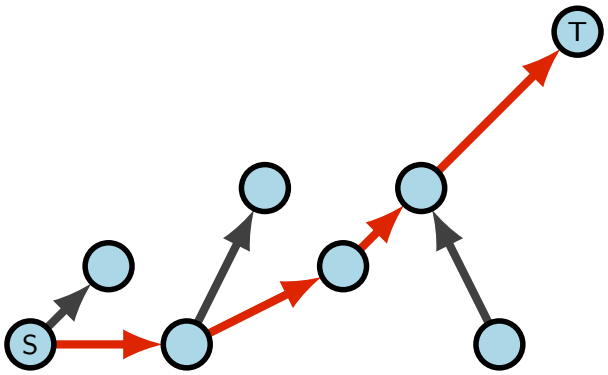



Figure 1: Sketch of dynamic graphs' snapshot with lateral movement from S to T

## Time aggregated dynamic graph $\mathcal{G}_{I_i} = (\mathcal{V}_{I_i}, \mathcal{E}_{I_i})$

**Given**

Marked point process: $(\mathcal{T}, \mathcal{X}, \mathcal{Y}) = \{(T_i, X_i, Y_i)\}_{i \geq 1}$

**Define**

Time interval: $I_i = \{t : t \in [(i-1)\delta, i\delta], \delta > 0\}, i \in \mathbb{Z}^+$

Node set: $\mathcal{V}_{I_i} = \{x_j : \exists t \in I_i \text{ for which } x_j \text{ exists}\}$

Edge set: $\mathcal{E}_{I_i} = \{(x_j, y_k) \in \mathcal{V}_{I_i} \times \mathcal{V}_{I_i} : N_{I_i}(x_j, y_k) > 0\}$

# Introduction. Dynamic graphs modelling and Contributions

- There are two main **modelling approaches** for dynamic graphs:
- **Global** [1, 13, 3, 9, 12, 11]: uncover a latent space ✓ **Information sharing**
- **Local** [17, 8, 14, 7, 10]: model activity at the node or edge level ✓ **Faster computation**

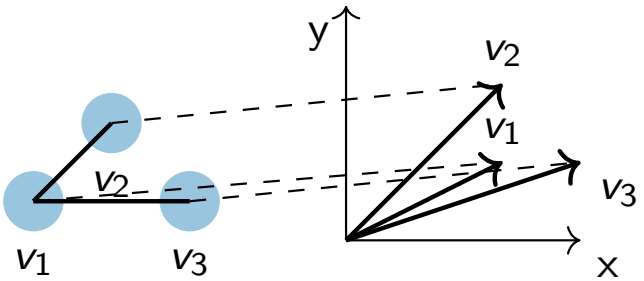



Figure 2: Sketch of global approach

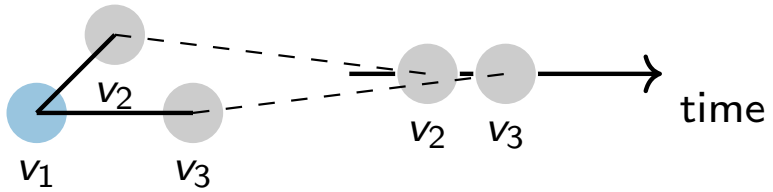



Figure 3: Local method example: node $v_1$ point of view

In this work:

- First we look at a local method, as these are better scalable; in particular, we extend a Bayesian non-parametric approach [8, 14] by **dropping the exchangeability** assumption;

$$P(\bullet, \bullet, \bullet, \bullet, \bullet, \bullet, \bullet, \bullet) = P(\bullet, \bullet, \bullet, \bullet, \bullet, \bullet, \bullet, \bullet),$$

- Secondly, we develop an algorithm, with more of a global flavour, to try and **reduce false alarms**, but which allows fast inference.

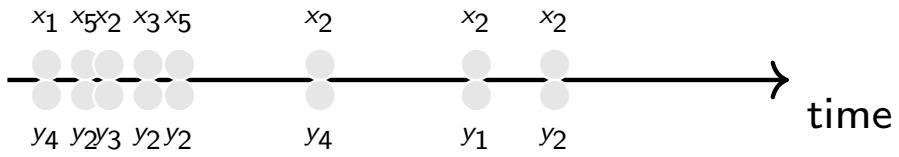
Imperial College London

# Outline

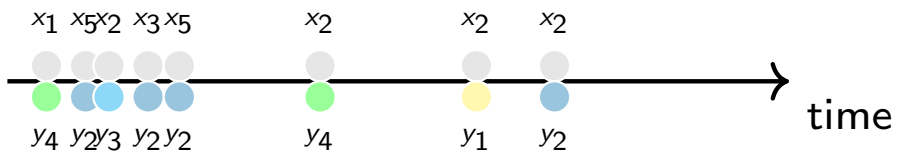
Imperial College London

# Methods I. Modelling network data with stochastic processes

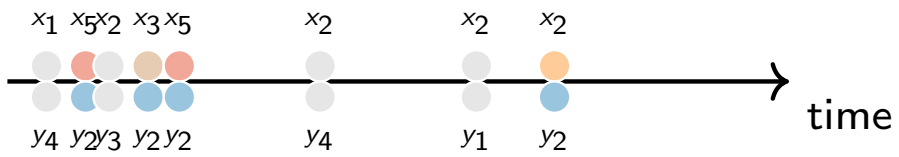
- Network authentication data as **marked point process** $(\mathcal{T}, \mathcal{X}, \mathcal{Y}) = \{(T_i, X_i, Y_i)\}$, $i \geq 1$, time $T_i \in \mathbb{R}^+$, marks $(X_i, Y_i) \in \mathcal{V} \times \mathcal{V}$, $\mathcal{V}$ nodes set;




- We decompose $p(x, y) = p(y)p(x|y)$ [8, 14], hence model $\mathbb{P}_{\mathcal{Y}} \overset{d}{\sim} \mathcal{BNP}(\theta_0, \mathbb{P}_0)$;




- Then model $\mathbb{P}_{\mathcal{X}|y_i} \overset{d}{\sim} \mathcal{BNP}(\theta_{y_i}, \mathbb{P}_0)$, $\mathcal{BNP}$ Bayesian non-parametric process;




- Calculate $\mathbf{p}(y_i)$, $\mathbf{p}(x_i|y_i)$ for a new observation;
- Approximate the link p-value $\mathbf{p}(x_i, y_i)$ using Fisher's method;
- Finally, link probabilities are combined to obtain **source nodes p-values**.

# Methods I. Literature approach

- [8, 14] use Dirichlet and Pitman-Yor processes (Dirichlet special case $d = 0$):

## Pitman-Yor Process

$n$ events, $K_n$ unique $\{x_j^*\}_{j=1}^{K_n}$. Pitman-Yor process if predictive distribution:

$$p(x|\mathbf{x}_n) \propto \begin{cases} \alpha + dK_n & x \neq x_j^* \ \forall j \\ N_n(x) - d & x \in \{x_j^*\}_{j=1}^{K_n} \end{cases}$$

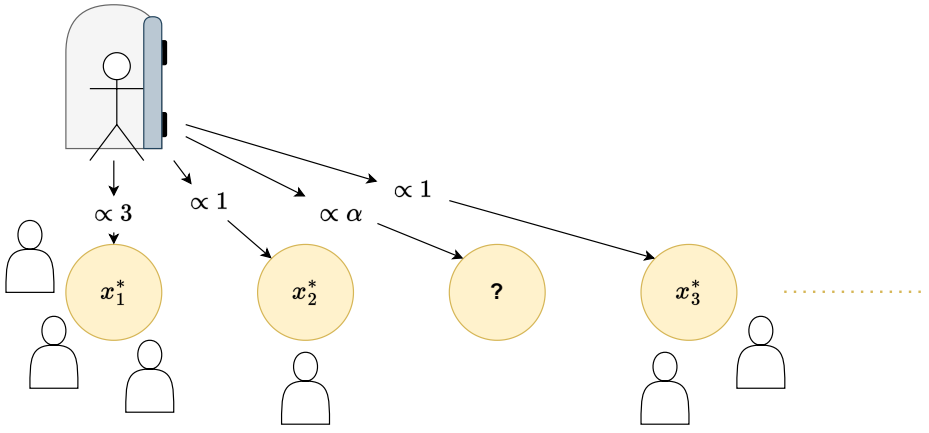



Figure 4: Restaurant metaphor for the Dirichlet process

- These processes have the exchangeability property:

## Exchangeability

$P(x_1, \cdots, x_n) = P\left(x_{\pi(1)}, \cdots, x_{\pi(n)}\right)$, where $\pi(\cdot)$ is a permutation of the events. Visually: $P($●, ●, ●, ●, ●, ●, ●, ●$) = P($●, ●, ●, ●, ●, ●, ●, ●$)$

# Methods I. Dropping exchangeability assumption

In this study we **drop the exchangeability assumption**, likely to not be verified in practice, by including dependency on inter-arrival times in three different ways:

1. Modelling of time process as Poisson process;
2. Streaming version of the Pitman-Yor process;
3. Explicit dependency on inter-arrival times with distance dependent Chinese restaurant process.

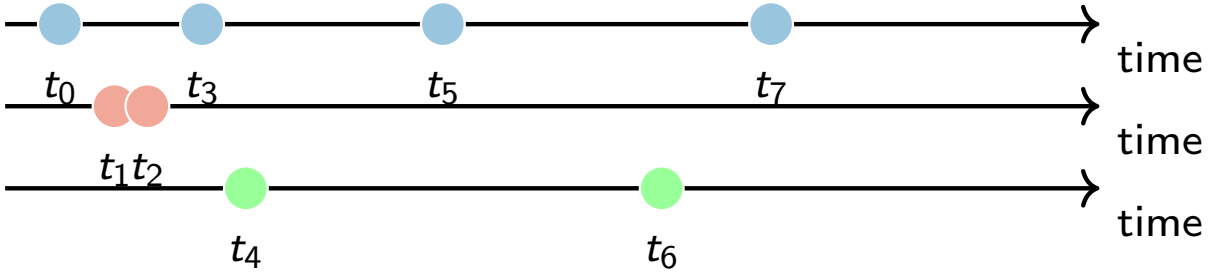



Figure 5: Modelling Poisson point process at link level

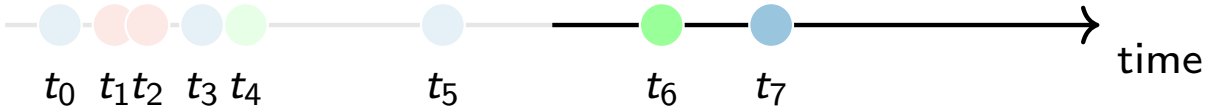



Figure 6: Streaming version of the process

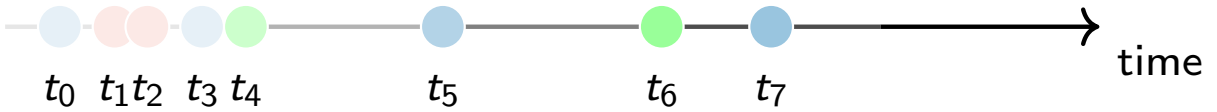



Figure 7: Distance-dependent CRP version

# Methods I. Poisson point process

- Consider $p(x, y, t) = p(y)p(x|y)\mathbf{p}(\mathbf{t}|\mathbf{x}, \mathbf{y})$ same process for $x$, $y$ as before;
- Need to guarantee congruency between the Pitman-Yor ordering and inter-arrival times, done using the **partitioning property** of the Poisson process [16]:

## Partitioned Poisson process

If the global number of links $N \sim Pois(\lambda)$, and each event at time $t$ has destination $y$ with probability $p(y|\mathcal{H}_y(t))$, then each sequence conditioned upon its category follows a Poisson process with:

$$N_y(t) \sim Pois\left(\lambda \int_0^t p(y|\mathcal{H}_y(s))\, ds\right).$$

Taking a step further, at link level:

$$N_{xy}(t) \sim Pois\left(\lambda \int_0^t p(x, y|\mathcal{H}_x(s), \mathcal{H}_y(s))\, ds\right).$$

✓ This formulation allows to calculate the p-value considering arrival times.
✗ Computations are slow, the integrand updates at any observation.

# Methods I. Streaming Pitman-Yor

## Streaming Pitman-Yor

Given the marked point process $\{\mathcal{T}, \mathcal{X}\}$ with realisations $\{(t_i, x_i)\}_{i=1}^{n}$, we seek to calculate the probability of a new observation $(t, x)$ as:

$$p(x|\{x_i : t - t_i < \delta\}).$$

✓ Forgetting of older observations.
✗ Abrupt forgetting and effect of the time-window choice.

# Methods I. Distance-dependent Chinese restaurant process

The distance-dependent Chinese restaurant process DDCRP [2] considers explicit dependency on inter-arrival times.

✓ Forgetting of older observations informed of inter-arrival.

✗ No significant downsides.

## Distance-dependent Chinese restaurant process

$x_k^*$ the group of customers to which customer $x_i$ is assigned, with $d_{ij}$ the distance between customer $i$ and $j$ and let $f(\cdot)$ be a decay function:

$$p(x_i = x_k^* | \mathbf{d}, \alpha) \propto \begin{cases} \alpha & x_k^* \neq x_j \forall j \\ f(d_{ij}) & x_k^* = x_j \end{cases} \quad (1)$$

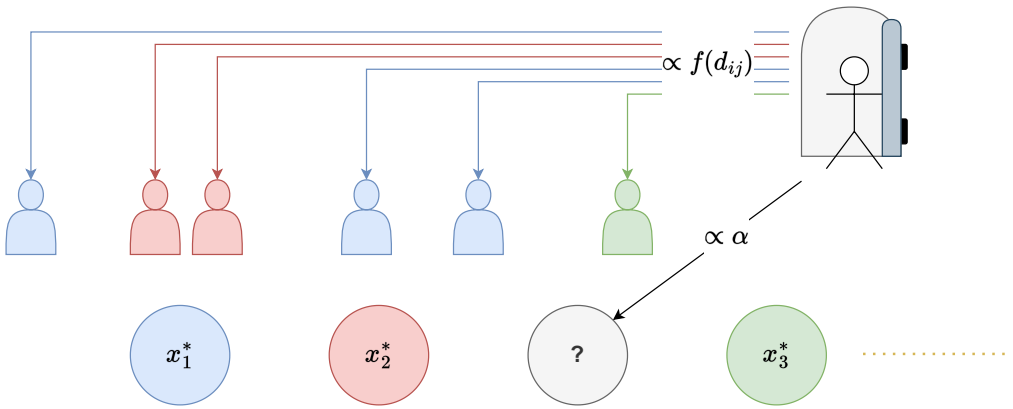



Figure 8: Restaurant metaphor for DDCRP

Imperial College London

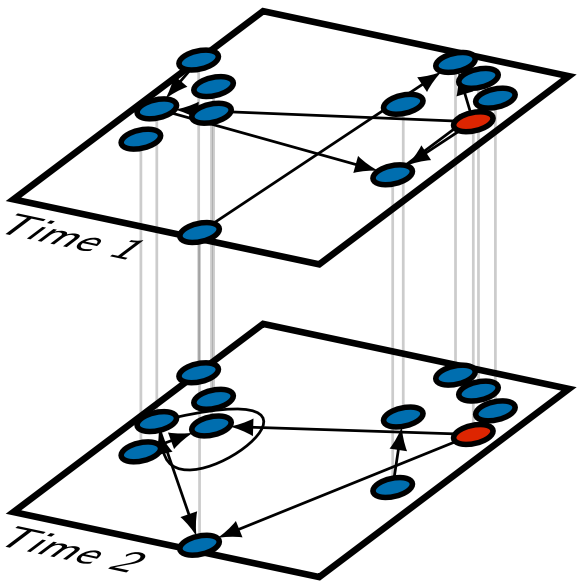
# Outline

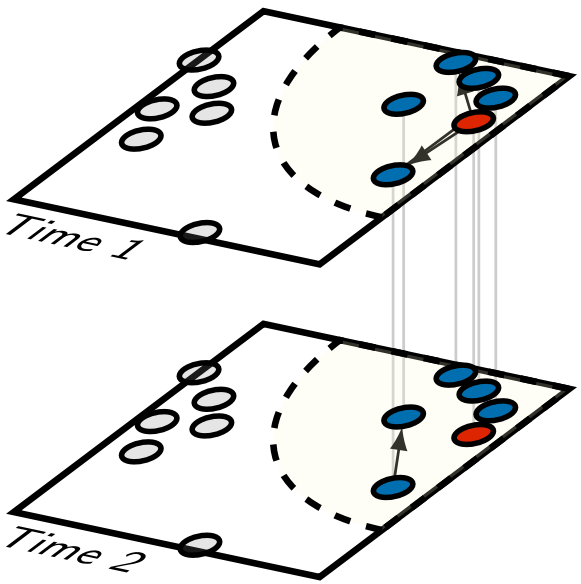
Imperial College
London

# Methods II. Globally informed anomaly detection

In this study we potentiate the first part of anomaly detection by verifying **if communication pattern in the neighbourhood of a potentially anomalous node are indeed unlikely within the entire network**. This requires the followings:

- Definition of a procedure of construction of the dynamic graphs;
- Definition of a model architecture to train on anomaly-free graphs;
- Algorithm to calculate an anomaly score, given the model;
- Algorithm to combine the results with anomaly scores from the first part of anomaly detection.

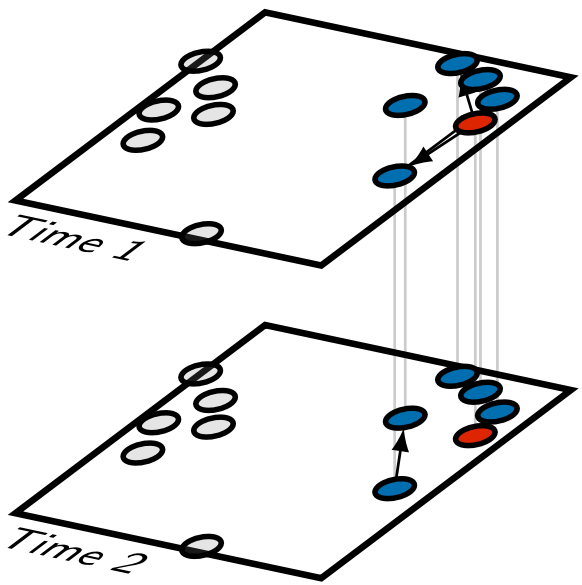


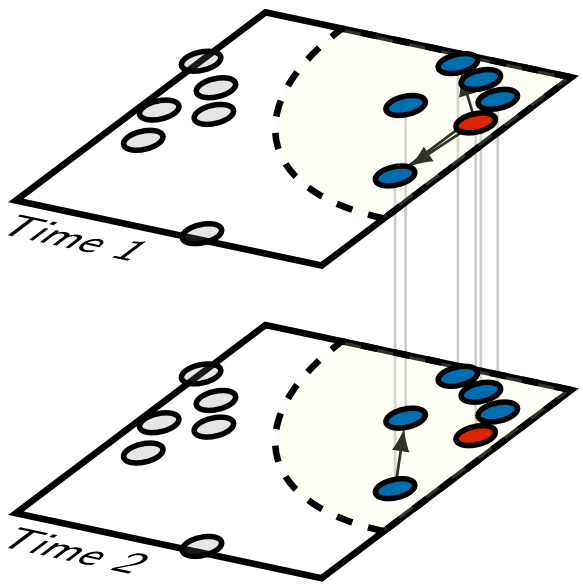



Imperial College London

# Methods II. Dynamic graph construction

- Given a node $x_i$ at time $t_i$, build dynamic graph $\mathcal{G}_T^{(x_i,t_i)}$;
- Time window: $T \in \mathbb{N}$ time aggregations with span $\delta \in \mathbb{R}^+$, the last of which is centred, in time, at $t_i$;
- Node set: set of nodes reachable from $x_i$ with an **out-directed** $k$-**steps path** in the time interval above;
- The edge sets, $\left(\mathcal{E}_1^{(x_i,t_i)}, \cdots, \mathcal{E}_T^{(x_i,t_i)}\right)$, defined as usual;



Imperial College London

# Methods II. Dynamic graph VAE

**Dynamic graph variational auto-encoder** [6], extended to consider directed graphs and a Gaussian mixture prior for extra flexibility.

✓ Flexible, adapts to different dimensions: we learn **node embeddings**

✓ Likelihood-based model, we learn a **probabilistic embedding**: allows to calculate the likelihood of unseen observations given the model

✓ The model predicts the adjacency matrix's entries as i.i.d. Bernoulli RV

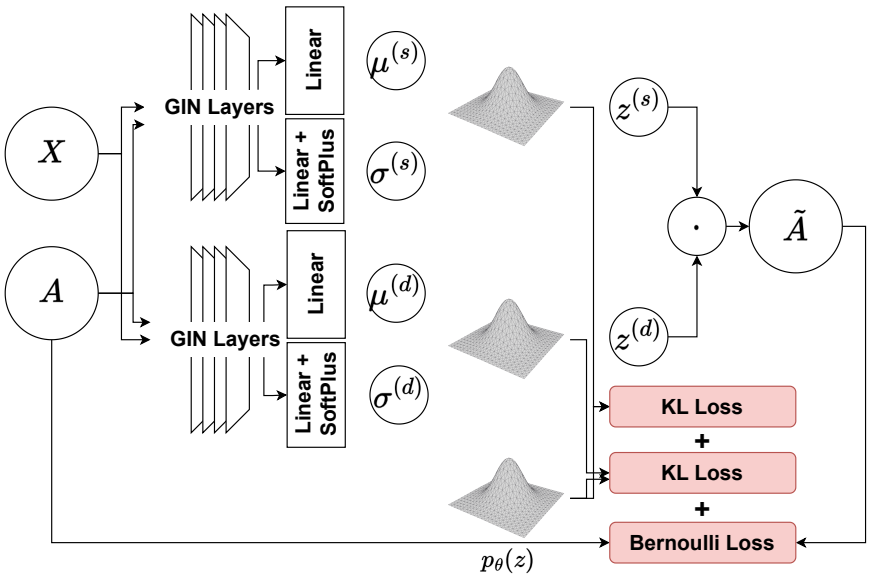



Figure 9: Dynamic graph variational auto-encoder DynVAE

$$A_{ij}|Z^s, Z^d \overset{iid}{\sim} Bernoulli(\tilde{A}_{ij})$$

$$\tilde{A}_{ij} = p_\theta(A_{ij} = 1|z_i^s, z_j^d) = SIGMOID({z_i^s}^T \cdot z_i^d)$$

After the model is trained, in inference we remove the dependency upon the latent space, to calculate the matrix $p(A)$, by integrating over it with Monte Carlo methods and **calculate a p-value**.

Imperial College London

# Outline

# Datasets

## Synthetic datasets with $|\mathcal{V}| = 100$

Creation of two **marked point process** $(\mathcal{T}, \mathcal{X}, \mathcal{Y}) = \{(T_i, X_i, Y_i)\}_{i \geq 1}$ with:

$$\mathcal{T} \sim \mathcal{PP}(\lambda(t)) \ , \ \mathbb{P}_{\mathcal{Y}} \overset{d}{\sim} PY(\alpha_0, d_0, \mathbb{P}_0)$$

and:

$$\mathbb{P}_{\mathcal{X}|y_i} \overset{d}{\sim} PY(\alpha_{y_i}, d_{y_i}, \mathbb{P}_0), \text{ or } \mathbb{P}_{\mathcal{X}|y_i} \overset{d}{\sim} DDCRP(\alpha_{y_i}, \beta, \mathbb{P}_0).$$

A **simulated intrusion is also included in the dataset**: a series of random walk paths is set out from the first hacked machine to the target machine.

## Los Alamos National Laboratory dataset [18] with $|\mathcal{V}| > 12000$

Consists of 58 consecutive days of Windows-based authentication events, with $1s$ time resolution, and is openly available at the following link.

Imperial College London

# Outline

Imperial College London

# Methods I results. Synthetic datasets

## Results evaluation

- Number of truly anomalous nodes ranked in the top 10 or top 20;
- Overall improvement of anomaly ranking with respect to the dataset without anomaly, or one of the methods taken as reference (Wilcoxon test, better ranking alternative hypothesis).

| Dataset 1 | PY | DP | DDCRP | Stream PY | Poisson and PY |
|---|---|---|---|---|---|
| N. of anomalies in top 10 | 6 | 5 | 6 | 9 | 7 |
| N. of anomalies in top 20 | 10 | 8 | 11 | 13 | 11 |

Table 1: Number of true anomalous source nodes in top 10 and top 20 for dataset 1

| Dataset 1 | PY | DP | DDCRP | Stream PY | Poisson and PY |
|---|---|---|---|---|---|
| Wilcoxon p vs. no anomaly | 0.67 | 0.88 | 0.94 | 0.86 | 0.89 |
| Wilcoxon p vs. reference | * | 0.95 | 0.21 | 0.10 | 0.79 |

Table 2: Wilcoxon test p-value calculated for ranking of all truly anomalous source nodes, against the same methodology on a dataset without anomaly and against other methods, for dataset 1

- For dataset I, **Stream PY** is the best performing model, **PY** and **DDCRP** follow; **PY** has the best overall ranking;
- For dataset II, **PY**, **DP** and **DDCRP** perform similarly;
- We can never reject the Wilcoxon null hypothesis firmly in this small dataset.

Imperial College London

# Methods I results. Los Alamos National Laboratory dataset

## Results evaluation

- Anomaly ranking of two, known, anomalous nodes: C17693 and C22409

| | PY | DP | DDCRP | STR. PY (12h) | STR. PY (36h) |
|---|---|---|---|---|---|
| C17693 | 332 | 304 | 81 | 1522 | 1182 |
| C22409 | 435 | 581 | 398 | 1610 | 1010 |

Table 5: Ranking of known anomalous source nodes according to model specification.

## Conclusions

- Methodology with Poisson process intractable due to long simulation times;
- DDCRP outperforms other methods;
- Streaming version of Pitman-Yor does not work well.

Imperial College London

# Outline

Imperial College London

# Methods II results. Sinthetic dataset

## Evaluation

- Tested three different aggregations: 5, 15 and 30 minutes;
- Tested only on dataset I, with anomaly scores from PY model in the first part of anomaly detection;

| | 1 | 1+2 (5min) | 1+2 (15min) | 1+2 (30min) |
|---|---|---|---|---|
| N. of anomalies in top 10 | 7 | 10 | 10 | 7 |
| N. of anomalies in top 20 | 11 | 12 | 12 | 11 |

| | 1 | 1+2 (5min) | 1+2 (15min) | 1+2 (30min) |
|---|---|---|---|---|
| Wilcoxon p vs. no anomaly | 0.968 | 0.964 | 0.933 | 0.963 |
| Wilcoxon p vs. reference | * | 0.083 | 0.999 | 0.942 |

Table 7: Result for dataset 1, obtained with hierarchical Pitman-Yor, using either only anomaly detection part 1 or 1 and 2 together, with different aggregation times. This table shows results obtained by combining **p-values** from part 1 and 2 using Fisher's method

- Improves upon sole first part of anomaly detection with 5 and 15 minutes aggregations;
- Results depend on aggregation times.

Imperial College London

# Methods II results: Los Alamos National Laboratory dataset

| | PY |
|---|---|
| C17693 | 332 |
| C22409 | 435 |

Table 8: Ranking of known anomalous source nodes according to model specification.

| | Part 1 p-value | Part 2 p-value | Combined | Ranking |
|---|---|---|---|---|
| C17693 | 0.0026 | 0.0000 | 0.0000 | 19 |
| C22409 | 0.0036 | 1.0000 | 0.0238 | 519 |

Table 9: Ranking of known anomalous source nodes using a Pitman-Yor hierarchical process for the first part of anomaly detection and 30 minutes aggregation time for the second part of anomaly detection, expressing the result of the DynVAE anomaly detection with p-values as these performed best on the synthetic dataset.

## Conclusions

- Difficult to conclude on performance for the second part of anomaly detection;
- Model training was terminated prematurely due to time constraints: there is scope for improvement, also by scanning hyper-parameters;
- The p-value distribution is peaked at 0 and 1.

Imperial College London

# Outline

# Conclusions. First part of anomaly detection

- Extended previous work on local anomaly detection in computer networks, based on Dirichlet and Pitman-Yor processes, by **dropping the exchangeability assumption in three different ways**: proposed three alternatives;
- Methods were compared and contrasted on two synthetic datasets with simulated intrusion, and on a real dataset;
- The **distance dependent Chinese restaurant process formulation of the problem was successful**, and a good challenger of state-of-art methods;
- The modelling of time processes at link level proved incompatible with computation time requirements in cybersecurity;
- The Pitman-Yor streaming version is quite poor, which highlights the challenge of defining a meaningful streaming window $\rightarrow$ forgetting factors version?

# Conclusions. Second part of anomaly detection

- Reduce false alarms, by verifying if the communication patterns in the neighbourhood of a potentially anomalous node are unlikely with respect to the type of activity seen in the entire network;
- Built dynamic sub-graphs in the neighbourhood of the potentially anomalous node, and for each node in the training set;
- Trained a **dynamic graph variational auto-encoder** to learn common pattern within the entire network;
- Deployed the model to calculate potentially anomalous graphs p-values given the model;
- The **effectiveness of the proposed approach was demonstrated on the synthetic dataset**, especially with 5 and 10 minute aggregations;
- On the Los Alamos National Laboratory dataset it is more difficult to conclude on performance, but there is scope to improve with longer model training and hyper-parameter tuning;
- One **challenge with the proposed approach is in the choice of the aggregation interval to construct dynamic graphs**, a choice that in practice might require to train different models for a set of different aggregations and use them altogether.

# References I


A. Athreya, D. Fishkind, K. Levin, V. Lyzinski, Y. Park, Y. Qin, D. Sussman, M. Tang, J. Vogelstein, and C. Priebe.
Statistical inference on random dot product graphs: A survey.
*Journal of Machine Learning Research*, 18, 09 2017.


P. I. Frazier D. M. Blei.
Distance dependent chinese restaurant processes.
*Journal of Machine Learning Research*, 12(2011):2461–2488, 2011.


I Gallagher, A. Jones, and P. Rubin-Delanchy.
Spectral embedding for dynamic networks with stability guarantees.
*ArXiv*, abs/2106.01282, 2021.


B. Gold, D. Curwin, and C. McClister.
Understand and investigate Lateral Movement Paths (LMPs) with Microsoft Defender for Identity.
`https://learn.microsoft.com/en-us/defender-for-identity/understand-lateral-movement-paths`, 2023.
Last accessed 23/08/01 at 09:47.

# References II


B. Gold, D. Curwin, C. McClister, S. Sagir, R. Wiselman, and M. Baldwin.
Lateral movement alerts.
https://learn.microsoft.com/en-us/defender-for-identity/lateral-movement-alerts, 2023.
Last accessed 23/08/01 at 09:48.


E. Hajiramezanali, A. Hasanzadeh, N. Duffield, K. R. Narayanan, M. Zhou, and X. Qian.
Variational graph recurrent neural networks.
*CoRR*, abs/1908.09710, 2019.


I. Hawryluk, H. Hoeltgebaum, C. Sodja, T. Lalicker, and J. Neil.
Peer-group Behaviour Analytics of Windows Authentications Events Using Hierarchical Bayesian Modelling.
*arXiv*, 2209.09769, 2022.


N. Heard and P. Rubin-Delanchy.
Network-wide anomaly detection via the dirichlet process.
In *2016 IEEE Conference on Intelligence and Security Informatics (ISI)*, pages 220–224, 2016.

# References III


P. D. Hoff, A. E. Raftery, and M. S. Handcock.
Latent space approaches to social network analysis.
*Journal of the American Statistical Association*, 97(460):1090–1098, 2002.


P. J. Laub, T. Taimre, and P. K. Pollett.
Hawkes processes.
*arXiv*, 1507.02822, 2015.
[math, PR].


F. Sanna Passino and N. A. Heard.
Mutually exciting point process graphs for modeling dynamic networks.
*Journal of Computational and Graphical Statistics*, 32(1):116–130, sep 2022.


F. Sanna Passino, Melissa J. M. Turcotte, and N. A. Heard.
Graph link prediction in computer networks using Poisson matrix factorisation.
*The Annals of Applied Statistics*, 16(3):1313 – 1332, 2022.


P. Rubin-Delanchy, J. Cape, M. Tang, and C. E. Priebe.
A statistical interpretation of spectral embedding: the generalised random dot product graph.
*arXiv*, 1709.05506, 2021.

# References IV


F. Sanna Passino and N. Heard.
Modelling dynamic network evolution as a Pitman-Yor process.
*Foundations of Data Science*, 3(1):293–306, 2019.


M. Schulze.
The schulze method of voting.
*CoRR*, abs/1804.02973, 2018.


K. Sigman.
*Batched Point Processes*.
School of Operations Research and Industrial Engineering, 2007.


M. Turcotte, N. Heard, and J. Neil.
Detecting localised anomalous behaviour in a computer network.
In H. Blockeel, M. van Leeuwen, and V. Vinciotti, editors, *Advances in Intelligent Data Analysis XIII*, pages 321–332, Cham, 2014. Springer International Publishing.


M. J. M. Turcotte, A. D. Kent, and C. Hash.
*Unified Host and Network Data Set*, chapter Chapter 1, pages 1–22.
World Scientific, nov 2018.

# References V


European Union.
Cybersecurity: main and emerging threats.
https://www.europarl.europa.eu/news/en/headlines/society/20220120STO21428/cybersecurity-main-and-emerging-threats, 2023.
Last accessed 21-03-2023 - 10:45.


European Union.
Cybersecurity: why reducing the cost of cyberattacks matters.
https://www.europarl.europa.eu/news/en/headlines/society/20211008STO14521/cybersecurity-why-reducing-the-cost-of-cyberattacks-matters, 2023.
Update 23-11-2022 - 14:29.